Estimating Local Decision-Making Behavior in Complex Evolutionary Systems

Zhenghui Sha
Graduate Research Assistant
School of Mechanical Engineering, Purdue University
West Lafayette, Indiana 47907
E-mail: zsha@purdue.edu

Jitesh H Panchal
Assistant Professor
School of Mechanical Engineering, Purdue University
West Lafayette, Indiana 47907
E-mail: panchal@purdue.edu

ABSTRACT

Research in systems engineering and design is increasingly focused on complex socio-technical systems whose structures are not directly controlled by the designers, but evolve endogenously as a result of decisions and behaviors of self-directed entities. Examples of such systems include smart electric grids, Internet, smart transportation networks, and open source product development communities. To influence the structure and performance of such systems, it is crucial to understand the local decisions that result in observed system structures. This paper presents three approaches to estimate the local behaviors and preferences in complex evolutionary systems, modeled as networks, from its structure at different time steps. The first approach is based on the generalized preferential attachment model of network evolution. In the second approach, statistical regression-based models are used to estimate the local decision making behaviors from consecutive snapshots of the system structure. In the third approach, the entities are modeled as rational decision-making agents who make linking decisions based on the maximization of their payoffs. Within the decision-centric framework, the multinomial logit choice model is adopted to estimate the preferences of decision-making nodes. The approaches are illustrated and compared using an example of the autonomous system (AS) level Internet. The approaches are generally applicable to a variety of complex systems that can be modeled as networks. The insights gained are expected to direct researchers in choosing the most applicable estimation approach to get the node-level behaviors in the context of different scenarios.

Keywords: Complex systems, decision making, multinomial logit choice models, evolutionary networks.

1 MOTIVATION FOR ESTIMATING LOCAL BEHAVIORS IN COMPLEX EVOLUTIONARY SYSTEMS

Research in engineering design and systems engineering has traditionally been focused on systems, such as automotive and aerospace systems, whose structure is under the direct control of designers. The design of such systems starts with the system-level requirements and is driven by top-down hierarchical decomposition, followed by the design for sub-systems and components. The component-level designs are integrated into a complete system and validated against system-level requirements. This general process is embodied in various systematic design methods (e.g., Pahl and Beitz [1]), systems engineering models (e.g., Systems Engineering Vee [2]), and systems engineering processes adopted by organizations such as NASA [3]. Due to the focus on such hierarchical design processes, the decision making literature within engineering design and systems engineering is primarily focused on decisions made by the designers during the design process.

There is an increasing importance of complex socio-technical systems whose designs are not directly controlled by the designers, but evolve as a result of decisions and behaviors of self-directed entities. An example of such a system is the smart electric grid, which consists of a wide range of decision-makers including consumers, utilities, micro-grid operators,

Corresponding Author: Jitesh H. Panchal Paper number: MD-13-1289

and the other participants of the distribution infrastructure. The energy producers, distributors, and utilities independently make technical decisions within rules and regulations to meet their objectives of system performance, reliability, security and load demand while maximizing their profits. The decisions made by the stakeholders affect the technical, social, economic, and environmental performance [4]. Other examples of such complex evolutionary systems include the Internet, air-transportation networks, and smart vehicle networks. We refer to such systems as "endogenously" evolving systems because their structures and properties are driven by the decisions made by entities within the system boundaries. In contrast, traditional hierarchical systems are exogenously designed by entities outside the system boundary. The key differences between traditional hierarchical systems and complex evolutionary systems are summarized in Table 1.

[Table 1 about here.]

Due to their fundamentally different nature, traditional top-down design approaches, discussed above, are not suitable for such complex evolutionary systems. From a design standpoint, the fundamental difference is that instead of directly controlling the system structure, the behaviors of the interacting entities must be modified in a bottom-up manner to achieve the desired system performance (such as robustness and resilience). Such modification of behavior can be achieved through the provision of incentives, imposition of penalties or taxes, and definition of rules. Therefore, *the role of design in the context of complex evolutionary systems is different*, thereby posing unique challenges from an engineering systems design standpoint.

A specific class of complex evolutionary systems consists of systems whose structure is modeled as endogenous networks in which the nodes[1] make local decisions about linking with other nodes. The underlying dynamics of complex endogenous networks can be understood by modeling the mappings across five levels, shown in Figure 1. The network structure (level 3) emerges from the node-level linking behavior (level 2), which is driven by the node-level preferences (level 1). The preferences of the nodes refer to the utility functions that the nodes maximize. The network structure in turn determines the network properties (level 4) and network performance (level 5). Consider the example of the Internet at an autonomous system (AS) level, where a node represents an AS and a link represents communication between two autonomous systems. The nodes make strategic decisions about linking with other autonomous systems in order to route data. These local decisions affect the global structure of the Internet. The global structure in turn affects its robustness and resilience to node failure (i.e., the performance). Thus, the node-level behavior is crucial to understanding the overall network performance.

[Fig. 1 about here.]

The process of traversing from the lower to the higher levels (i.e., 1 to 5) is *analysis*, in which the performance of the network is determined in terms of the node-level preferences and behaviors. On the other hand, achieving targeted performance by determining how to modify the node-level preferences can be viewed as a *design problem* [5]. To address the design problem in endogenous networks, it is crucial to understand the node-level preferences that result in observed network structures, and how the node-level preferences and behaviors influence the overall network structure. The focus in this paper, as shown in Figure 1, is on the estimation of node-level behaviors and preferences. Estimating the node-level preferences and behaviors in real-world networks can help in:

1. accurately modeling the evolutionary dynamics of endogenous networks, and
2. determining mechanisms for influencing the node-level behaviors and the provision of incentives to achieve targeted system performance.

The local behaviors in complex evolutionary systems can be estimated in two different ways. First, the local behaviors can be directly estimated by conducting surveys and interviews of the decision makers involved in making decisions. However, in many cases, the designer may not have direct access to the decision makers. In such cases, an alternate approach is to infer the decisions indirectly from the decision patterns from past data. In complex networks, this can be achieved by estimating local behaviors from the system's structure itself. In this paper, we present three approaches to estimate the node-level behaviors and preferences in complex endogenous networks from their structure at different time steps. The inputs are data about nodes and their connectivity at discrete timesteps, and the outputs are behaviors of the nodes. The approaches are discussed in Section 2 and illustrated in Section 3 using the AS-level Internet network, which is an example of complex evolutionary system. The approaches are used to estimate the AS-level linking behavior that results in the observed evolution of the Internet. A comparative analysis of the three approaches is presented. Closing thoughts are presented in Section 4.

2 APPROACHES FOR ESTIMATING NODE-LEVEL BEHAVIORS

[1] In network terminology, a network is composed of nodes and links. A network can be mathematically represented as a graph where nodes are represented as vertices and the links are represented as edges. Within complex networked systems, a node can refer to individual decision makers or other entities such as organizations that make decisions.

An overview of the three approaches proposed in this paper is presented in Table 2. The first approach is based on hypothesized node-level behaviors. The node-level behavior is considered to represent reality if the generated networks have structures similar to the real networks. These models are primarily used for explaining the evolution of real-world complex systems. In this paper, we adopt the **generalized preferential attachment (GPA)** [6] as the hypothesized node-level behavior model. The parameters in this model can thus be estimated based on the relationship between the network structure and the degree-based node-level behavior (probability of linking) which is derived from the continuum theory of network evolution presented by Albert et al. [7].

[Table 2 about here.]

In the second approach, the node-level behaviors are derived by analyzing how the networks change between two consecutive instances. The nodes and links created (or removed) between the two instances are extracted first, and the node-level behavior is deduced by using **statistical regression** techniques. In this paper, we propose to estimate the node-level behaviors with linear regression model.

In the third approach, network evolution is modeled as a decision-making process where the nodes are decision-making agents and their behaviors are based on utility maximization. While the first two approaches can help in revealing the node level behavior, the node-level preferences are not deduced from the network structure. The third approach can be used to determine both the behavior and the node-level preferences. We adopt the **multinomial logit choice model** to estimate the local entities' preferences and behaviors, but we develop the method to deal with large choice set so that the proposed approach can be applied to the large scale complex systems.

Our rationale for choosing GPA is that it is one of the widely used approaches for modeling complex networks. GPA has also been used for modeling the Internet topology and its properties. Statistical regression techniques are chosen for comparison purposes because they are widely used within the design literature for data-driven modeling in complex systems. We propose the use of decision-based models (third approach) to capture the node-level preferences, which are not currently modeled within the literature. To ensure that the comparison between the models is meaningful, we use the same input data in all approaches, and we only consider the structural aspects of the networks in the decision model. Details of the three approaches are presented next.

2.1 Approach 1: Generalized Preferential Attachment

The preferential attachment model for complex networks was initially proposed by Barabasi and Albert [6]. In this model, a new node preferentially links to existing nodes based on certain characteristics of the target node. Network evolution in this approach is assumed to follow two mechanisms: growth and preferential attachment [8]. The growth mechanism prescribes that at each time step, one new node is added with m edges linking the new node to m existing nodes in the network. In the simple preferential attachment model initially proposed by Barabasi and Albert [6], the probability of link creation between a new node and an existing node is linearly proportional to the degree of an existing node.

Preferential attachment has been widely accepted in the field of complex networks research and has been utilized for modeling real-world complex networks such as the Internet [9], the World Wide Web [10], and networks of metabolic reactions [11]. Existing literature [12] has shown that the degree-based preferential attachment mechanism has better performance in modeling real-world complex evolving networks with a power-law degree distribution. The degree-based linear preferential attachment model has been extended to generalized preferential attachment (GPA). An overview of GPA is presented next.

2.1.1 Overview of generalized preferential attachment (GPA)

In the GPA model, the affinity of a node to link with an existing target node j at time t is modeled as:

$$V_j(t) = G_j(t) d^{\tau}{}_j(t) + A_j(t) \tag{1}$$

where, the V, G, A and d are functions of the node j and time t. $G_j(t)$ is the fitness value of node j at time t, and $A_j(t)$ is the additional attractiveness of node j at time t. $d_j(t)$ stands for the degree of node j at time t, which is the number of neighbors of node j. Using Equation (1), the probability of an arbitrary node j getting chosen for connection among J nodes is equal to:

$$\begin{aligned} P_j &= \frac{V_j(t)}{\sum_{i=1}^{J} V_i(t)} \\ &= \frac{G_j(t)}{\sum_{i=1}^{J} V_i(t)} d_j(t) + \frac{A_j(t)}{\sum_{i=1}^{J} V_i(t)} \end{aligned} \tag{2}$$

This probability function is assumed to result in the evolution of the network between two time steps. Furthermore, it is assumed that i) the network is undirected, ii) the fitness value for all the nodes is the same and is time independent, thus $G_i(t)$ is constant, iii) the additional attractiveness for each node is time independent and a constant, thus $A_i(t)$ is constant, and iv) the affinity V in Equation (1) is a linear function of the node degree, i.e., $\tau = 1$. Therefore, Equation (1) can be modeled as:

$$V_j(t) = d_j(t) + A_j \tag{3}$$

These assumptions are made to enable direct comparison with the other two approaches. By relaxing these assumptions, detailed models with more parameters can be generated. For example, if a directed network is used, then two separate probabilities are needed to model the creation of incoming and outgoing links. If different fitness values are used for different nodes, an additional parameter is added for each node, which increases the data requirements for parameter estimation. Similarly, considering time varying additional attractiveness also adds additional parameters in the model. Hence, for the purpose of this comparative study, we decided to limit the number of parameters. In the future studies, we will investigate the effects of relaxing these assumptions.

If $G_j(t)$ is assumed to be the same for all nodes, its impact can be accounted for by scaling the additional attractiveness parameter as follows:

$$\begin{aligned} P_j &= \frac{V_j(t)}{\sum_{i=1}^{J} V_i(t)} = \frac{Gd_j + A_j}{\sum_{i=1}^{J} (Gd_i + A_i)} \\ &= \frac{d_j + \frac{A_j}{G}}{\sum_{i=1}^{J} \left(d_i + \frac{A_i}{G}\right)} \end{aligned} \tag{4}$$

The additional attractiveness (A) and the node's degree determine the complete behavior model of a linking node. Based on the prior work by Sha and Panchal [5], it has been shown that additional attractiveness (A) has a significant impact on the network structure and network performance. The additional attractiveness is estimated through the degree distribution function obtained by using the continuum theory of network evolution, discussed next.

For analyzing the evolutionary process in this model, the continuum theory approach proposed by Albert et al. [7] provides a bridge between the network structure and the node-level properties such as the degree. With the continuum theory, the effect of additional attractiveness (A) on the structure, specifically the degree distribution of the resulting network, can be analyzed. According to the growth mechanism described above and the model proposed in Equation (3), the changing rate of a node j's degree d_j is given by:

$$\frac{\partial d_j}{\partial t} = m \frac{d_j + A}{\sum_{i=1}^{J-1} (d_i + A)} \tag{5}$$

where m is the number of edges linking to a new node in each timestep. Following the steps in [6], as the network becomes large,

$$\lim_{t \to \infty} P[d_j(t) \geq d] = \left(\frac{d + A}{m + A}\right)^{-\gamma} \tag{6}$$

Thus, as $t \to \infty$, the asymptotic complementary cumulative degree distribution (CCD) has the form:

$$F(d) = P[d_j(t) \geq d] \propto d^{-\gamma} \tag{7}$$

where,

$$\gamma = f(m, A) = \left(2 + \frac{A}{m}\right) \tag{8}$$

Equation 8 indicates that different degree distributions are associated with different A values. Hence, the A value can be used to differentiate the network structures. The degree distributions generated for representative values of A are illustrated

in Figure 2. By fitting the power-law degree distribution, we can determine the exponent γ using regression techniques to deduce the values of additional attractiveness, A, which defines the node-level behavior (Equation (4)). In Section 2.1.2, we introduce the techniques used for fitting the power law degree distribution.

[Fig. 2 about here.]

2.1.2 Fitting the Power-law Degree Distribution

A simple approach for fitting the power law degree distribution is the ordinary least square (OLS) regression [13]. The power law distribution in Equation (6) follows a straight line on a double logarithmic plot. Therefore, a commonly adopted technique to estimate the power law behavior in empirical data is to measure the frequency of nodes with degree d in the network and to plot such frequency on the double logarithmic axis. Then, a linear model:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{9}$$

can be used where β_0 is the intercept, β_1 is the slope and ε is the random error in the observation. The OLS regression can be utilized to fit the power law degree distribution with variable x equal to $ln(d)$ and observation value y equal to $ln(P)$. The estimation on the parameter β_1, which is the slope, γ, is the exponent in the power law.

In practice, the power law often applies only for values greater than some minimum value x_{min}. In such cases, the OLS regression method can produce inaccurate estimates of the parameters for power law distributions especially for the "tail" of the distribution where the values are under x_{min}. To address this issue Clauset et al. [14] proposed an effective statistical framework for fitting the power law distribution to empirical data. The approach combines maximum likelihood fitting with goodness-of-fit tests based on the Kolmogorov-Smirnov (KS) statistic and likelihood ratio [15]. The key idea for estimating the exponent, γ, correctly is to first identify the lower bound x_{min} of power law behavior in the data. Hence, the parameter x_{min} is first chosen, and then γ of the power law is fit using maximum likelihood estimation. Then, with the estimated x_{min} and γ in the first step, the power law hypothesis is tested by calculating the p-value for goodness-of-fit test that quantifies the plausibility of the hypothesis. A power law hypothesis is considered plausible for the data if the resulting p-value is greater than 0.1. Finally, the power-law models derived using alternate values of x_{min} are compared via a likelihood ratio test [16]. If the calculated likelihood ratio is significantly different from zero, then its sign indicates whether an alternative is favored or not.

Once we have the estimation on the exponent γ in the power law, we can obtain the additional attractiveness (A) using Equation (8). The m values for different networks can be obtained by plotting the number of nodes (J) versus the number of edges (E) in the network over time. An OLS regression between the number of new nodes and the number of new links can be used to estimate m. An illustrative example is presented in Section 3.3.1.

2.2 Approach 2: Statistical Regression-based Model

In the second approach, the linking behavior is determined by comparing two consecutive instances of the network structure. The node-level behavior is then obtained by fitting the node-level linking probability data using regression techniques. Consider a complex endogenous network that evolves from network $N(t_0)$ to network $N(t_1)$ during an interval $[t_0\ t_1]$. In order to obtain the behavior of the added nodes, we calculate each target node's probability of getting a connection from the newly added nodes.

From the datasets $N(t_0)$ and $N(t_1)$, we obtain the number of new nodes entering the network during $[t_0\ t_1]$. For each newly added node, the target nodes are identified. Based on the network structure from the dataset $N(t_0)$, the degrees of these target nodes are extracted. In the following step, all nodes in $N(t_0)$ are divided into different groups based on their degrees. All nodes with degree d are grouped into a group S_d. The number of nodes within a group S_d is represented as n_d. If the number of new links created with existing nodes in S_d is denoted by l_d, and the total number of links created during the interval $[t_0\ t_1]$ is L, then the probability of a group S_d receiving a link is:

$$P(S_d) = \frac{l_d}{L} \tag{10}$$

This is based on the assumption that each linking decision made by a node is a mutually exclusive event, the probability of all nodes getting a connection in the same group is the sum of the probability of each node in this group getting a connection. Considering all nodes to be identical, an individual node with degree d has the following probability of receiving a connection:

$$P(d) = \frac{1}{n_d} P(S_d) \tag{11}$$

Once the probability of an individual node with degree d getting connections has been determined, the degree versus probability relationship is plotted. By using an appropriate fitting model using OLS regression, the node-level behavior can be obtained. The application of the proposed approach for the case study is presented in Section 3.3.2.

2.3 Approach 3: Multinomial Logit Choice Model

In the third approach, we model the network evolution using a decision-making framework. Here, each new node is considered as a decision maker that maximizes its own utility function, u. Say a node i is a decision maker at time t, its decision on which target node j to link to is based on the maximization of u_i based on the network topology at time t. At a given time, node i has J alternatives to choose from. The decision-maker node i selects a target node j for creation of an edge based on the utility function. The utility function, u_i, can depend on the structural parameters of the nodes (e.g., degrees), or non-structural factors (e.g., capacity, cost, etc.).

We use the random utility discrete choice models (DCM) to estimate the utility functions that the nodes maximize while selecting other nodes to link to. Specifically, we use the multinomial logit choice model to model the decisions of the nodes. A brief introduction to DCM and multinomial logit is provided in Section 2.3.1. Modeling the network evolution using the multinomial logit choice model is discussed in Section 2.3.2.

2.3.1 Discrete Choice Model and Multinomial Logit

Multinomial logit is a technique for discrete choice analysis (DCA) [17] in which the size of the choice set, J, for the decision maker is greater than two. DCA has been widely used to model and forecast product demand by capturing individual customers' choice behavior, especially for demand estimation [18]. Earlier applications were in the field of transportation engineering, but later, DCA was extended to the field of product design to model consumer preferences under uncertainty [19]. DCA is based on the assumption that individuals seek to maximize their personal utility, u, when selecting an alternative from the choice set. The decision maker knows the utility function and uses it for making decisions. However, the observer is only able to observe the choices made by the decision maker. Therefore, from the observer's perspective, the utility function is random. DCA assumes that the individual's utility is a sum of two components [20]:

1. *Systematic component*, denoted by V_j, which is a function of different observed attributes x_j of the alternatives which can be either alternative specific or decision-maker specific [21]. It is assumed that this component is a linear combination of the observed attributes: $V_j = \beta_j{}^T \cdot x_j$ where β_j are the parameters corresponding to the observed attributes x_j. The observed attributes can also be referred to as the explanatory variables that describe the decision maker's utility function. It is important to note that this component is deterministic from the observer's point of view.
2. *Unobserved component*, ε_j, which can be represented as a random variable from the observer's point of view. This error term includes the impact of all unobserved variables that affect the utility of a specific alternative. Thus,

$$u_j = \beta_j^T \cdot x_j + \varepsilon_j \tag{12}$$

Based on utility maximization, the probability that alternative 1 is chosen by the decision-maker n from a choice set containing two alternatives, is equal to the probability that the utility of alternative 1 is greater than the utility of alternative 2. This can also be represented as:

$$\begin{aligned} P_n(1|[1,2]) &= P_n(u_{1n} > u_{2n}) \\ &= P_n(V_{1n} + \varepsilon_{1n} > V_{2n} + \varepsilon_{2n}) \\ &= P_n(\varepsilon_{1n} - \varepsilon_{2n} > V_{2n} - V_{1n}) \end{aligned} \tag{13}$$

In order to predict the choice probability, methods such as binary logit, probit, multinomial logit and mixed logit [20] can be used. The primary difference between these models is the assumption about the probability distribution of the unobserved component. In this paper, we use the multinomial logit model where the error terms ε_j are assumed to be independent and identically distributed across choice alternatives and observations (decision-maker), and follow a Gumbel distribution [17]. Then, the probability that decision-maker n chooses alternative 1 over alternative 2 is:

$$P_n(1|[1,2]) = P_n(u_{1n} > u_{2n}) = \frac{e^{V_{1n}}}{e^{V_{1n}} + e^{V_{2n}}} \tag{14}$$

The binary alternatives scenario has been extended to the Multinomial Logit (MNL) model, see Equation (15), that describes the probability of alternative j being chosen by decision-maker n from among a choice set containing J alternatives.

$$P(j|C_J) = \frac{e^{V_{jn}}}{\sum_{i=1}^{J} e^{V_{in}}} \tag{15}$$

Estimation techniques such as the maximum likelihood and Bayesian estimation can be used to determine the coefficients β in Equation (12) such that the model's prediction of choices matches the observed choices as closely as possible. In practice, existing statistical analysis software can be used for estimation of parameters in a multinomial logit model. In this paper, the mlogit package [22] for R [23] is used.

2.3.2 Describing the Network Evolution using the Multinomial Logit Model

If the complex network evolution is based on node-level decision-making process, the principles from discrete choice theory can be utilized to estimate the utility function, and the resulting choice probability (i.e., the probability of a newly added node in $N(t_1)$ choosing an existing node from network $N(t_0)$). The choice probabilities of individual nodes can then be aggregated to estimate the aggregate node-level behaviors.

In the multinomial logit choice model, the observations from a researcher's point of view are the newly added nodes who choose a target node to link to. The alternatives are the existing nodes during the previous time step. For the selection of each node's utility, if the observed variable x in the systematic component is only alternative specific, e.g., the node's degree, then the systematic component is:

$$V_j = \beta_{0j} + \beta_{1j} d_j \tag{16}$$

This corresponds to Equation (1) in which β_{0j} stands for the additional attractiveness and β_{1j} is the node fitness. The resulting probability of the node alternative j that gets connection from the decision-making node n is given by Equation (15). Note that this is fundamentally different from Equation (4). Finally, the parameters β_{0j} and β_{1j} can be estimated using maximum likelihood estimation techniques.

For large sized networks, the choice set may be large (e.g., over 10,000 nodes). To reduce the complexity of parameter estimation, the size of the choice set can be reduced by grouping the nodes with same degree together as one alternative. The resulting probability is the one that a group is chosen over other groups by a newly added node. The probability of connecting to a node within a group can then be obtained by randomly choosing a node from the group to which that individual node belongs.

The utility function can be further refined by considering other structural and non-structural parameters of the network. By considering more attributes of the alternatives, such as, the clustering coefficient [24], and betweenness centrality [25] different hypotheses about the utility functions can be generated and tested. Through this approach, accurate models of choices that match the observed choices can be obtained, and the factors (besides node's degree) that constitute the additional attractiveness of a node can be investigated. Hence, this approach is richer than the two approaches described in Sections 2.1 and 2.2.

3 ILLUSTRATIVE EXAMPLE: AS LEVEL INTERNET TOPOLOGY

In this section, the approaches discussed in Section 2 are applied to the autonomous system (AS) level Internet network. While the approaches are applicable to a variety of complex networked systems, the Internet is chosen as an example because of the availability of data. The goal here is to illustrate how these approaches can be implemented in practice to deduce the AS-level behaviors and decisions that result in the observed evolution of the Internet. As discussed in Section 1, the Internet is an ideal example of a complex evolving system that has emerged based on the decisions made by independent decision making entities. Estimation of local decision-making behaviors of the entities is important for understanding how the structure of the network evolves. The knowledge of the local behaviors can help in providing incentives to the autonomous systems to direct their linking behavior towards structures with desired performance characteristics such as robustness and resilience. A brief introduction to the AS-level Internet is provided in Section 3.1. The dataset of the AS-level Internet network is described in Section 3.2. The results from the different approaches are presented in Section 3.3. Finally, a discussion of the results is presented in Section 3.4.

3.1 Introduction to AS-level Internet

The Internet is a network of interconnected computers consisting of private, public, academic, business, and government networks linked by various networking technologies. The Internet network can be treated as an endogenously evolving

complex network because of the decentralized governance in usage and access policies. The topology of the Internet can be studied at three different levels [26]:

1. *IP level*, which is composed of the interfaces of routers that exchange information because each interface owns an IP on the Internet.
2. *Router level*, which is the interconnection of routers on the Internet. It represents cables, satellite or radio links, etc. This physical infrastructure is the one over which information is routed.
3. *AS level*, which models the way autonomous systems are interconnected. The Internet can be divided into thousands of domains connected with each other. Each domain is a collection of hosts connected via routers and switching facilities.

An AS is defined as "a connected group of one or more IP prefixes run by one or more network operators which has a single and clearly defined routing policy" [27]. Examples of autonomous systems include ISPs, corporate networks, and universities. An ISP can have one or more autonomous systems. Autonomous systems are connected via dedicated links or public network access points. A link between two autonomous systems represents a contract to forward data to each other over the link. Each AS can choose its policy to select the best route for data based on commercial contractual relationships. These contracts and AS-level policies play a significant role in determining the structure of the Internet and its overall performance [28]. The AS-level topology also influences the definition of routing protocols such as the Border Gateway Protocol (BGP). Hence, it is an important and appropriate level of abstraction to model the decisions that result in the structure of the Internet.

3.2 Data source

Publicly available data sources are available for Internet network data. Skitter, Archipelago (Ark) from Cooperative Association for Internet Data Analysis (CAIDA) [29] and the RoutView [30] from the University of Oregon are the three main projects for collecting the Internet topology data at the AS level. Specifically, the Ark project is an upgraded version of the previous Skitter project operated by CAIDA after Skitter served nearly a decade and was retired on Feb. 8^{th}, 2008 [29].

The dataset adopted in this paper is from CAIDA AS Relationships Dataset from January 2004 to November 2007. There are 122 files in total, each file containing a full AS graph derived from a set of BGP table snapshots used to exchange routing information between ASes.

3.3 Estimating the AS-Level Behavior in Internet Network using the Three Approaches

In this section, we present the results from the three approaches, starting with the Generalized Preferential Attachment (GPA) approach.

3.3.1 Results from Approach 1: Generalized Preferential Attachment

The first step in this approach is to develop a fit for the degree distribution of the network. Figure 3 shows an example degree distribution for the AS-level Internet on Jan. 5^{th}, 2004, along with the OLS regression model. The figure shows that a power law [31] is a good fit for the degree distribution of the network. Since the degree distribution is plotted on double logarithmic axes, the slope of the fitting line is the exponent γ in the power law relation (see Equation (9)).

[Fig. 3 about here.]

To determine how the power law distribution changes with time, we extract the exponents of the degree distribution for all 122 snapshots of the network from 2004 to 2007. Since the network size increases monotonically over time, the exponent is plotted against the network size that corresponds to each network at each time in the x-axis. The exponents are plotted in Figure 4. It is observed that this exponent γ increases with the network size.

[Fig. 4 about here.]

Based on Equation (8), the additional attractiveness (A) in the node-level behavior model can be evaluated using the exponents (γ) and the number of new links added in each time step (m). The m-value can be identified by plotting the number of nodes (J) vs. the number of edges (E) as the network grows. This plot is shown in Figure 5. It is observed from the figure that the number of edges increases linearly with the number of nodes. The slope of the line shows that for each new node, about 2 new edges are added. This indicates that $m \approx 2$.

[Fig. 5 about here.]

The additional attractiveness, A, can be calculated using m and γ based on Equation 8. We obtain that the A-value increases from -1.78 to -1.73. One-tail test on the slope of the fitting function for the parameter γ (i.e., S_γ) is performed. The t-statistic corresponding to $\{H_0^1 : S_\gamma = 0$ vs. $H_1^1 : S_\gamma > 0\}$ is 21.34, resulting in the p-value < 0.001. Hence, we claim that the

slope of γ_1 is statistically significant. This indicates that as the Internet grows, the additional attractiveness in the network increases, which impacts the node's linking preference. The impact of additional attractiveness on the linking behavior is discussed in detail by Sha and Panchal [5]. As A increases, more nodes have the opportunity to be connected.

We also used the approach suggested by Clauset et al. [16] (discussed in Section 2.1.2) to fit the degree distribution using the maximum likelihood estimator. The resulting exponents for the 122 networks are shown in Figure 6. By performing the t-test on the slope in the figure, the p-value corresponding to $\{H_0 : S_\gamma = 0 \text{ vs. } H_1 : S_\gamma \neq 0\}$, where S_γ is the slope of the parameter γ, is 0.14. Hence, there is no statistically significant change in γ. Note that the exponents in the power-law shown in Figure 6 are also greater than those in Figure 4. This can be explained as follows. The main difference in this method is that a minimum bound value x_{min} is estimated beyond which the fit is close to power law, and the "tail" of the distribution with values of degree lower than x_{min} are not considered in the fitting. Therefore, the resulting power law curve is only for the part of the data that is regarded as a true power law. Since the change in degree for the nodes that have low degree in the network is not substantial, the fit for that part of data does not change significantly. Because of the positive linear relationship between the exponent γ and the A value, the A value in turn remains unchanged as the network grows.

[Fig. 6 about here.]

3.3.2 Results from Approach 2: Statistical Regression

In this section, we utilize the approach presented in Section 2.3 to the AS-level. Figure 7 shows the node-level linking behavior (i.e., probability of new node linking to an existing node) in three pairs of consecutive network snapshots:

a) Jan. 5^{th}, 2004 (N1) - Feb. 2^{th}, 2004 (N2),
b) Aug. 28^{th}, 2006 (N59) - Sep. 4^{th}, 2006 (N60), and
c) Nov. 5^{th}, 2007 (N120) - Nov. 12^{th}, 2007 (N121).

[Fig. 7 about here.]

The plot is shown on a log-log scale. We use the degrees of existing nodes based on the previous snapshot of the network structure. We fit the data with a power function $y = \alpha x^{\beta}$ where y is the probability of linking to a node, and x is the degree of the target node. Thus, the linking probability of the node j is:

$$P_j = \alpha {d_j}^{\beta} \tag{17}$$

The parameters α and β are estimated using OLS regression on $ln(d)$ vs $ln(P)$. Furthermore, as shown in the figure, the parameters of the three fitting functions are close to each other, which indicates that the node-level behavior is consistent over time. This conclusion about the node-level behavior is different from the one obtained using the first approach (see Figure 4). However, the result is in agreement with the fit using the maximum likelihood estimation (see Figure 6).

To further validate this conclusion, we extract the node-level behaviors from all the 121 changes in the network from Jan. 2004 and 2007, and then determine the parameters of the fit (α and β). Figures 8 and 9 show the two parameters for the 121 timesteps. We performed two separate hypothesis tests to detect whether there have been any statistical significant changes in α and β over the 121 evolutions. The p-value corresponding to $\{H_0^1 : S_\alpha = 0 \text{ vs. } H_1^1 : S_\alpha \neq 0\}$ and $\{H_0^2 : S_\beta = 0 \text{ vs. } H_1^2 : S_\beta \neq 0\}$ are 0.03 and 0.04 respectively. Here, S_α and S_β are the slopes corresponding to parameters α and β in Figures 8 and 9 respectively. Hence, we claim that there has been no statistically significant change in slopes of these two parameters at a 2% level of significance. Hence, we conclude that the node-level behavior of the ASes is consistent during 2004 and 2007. The average values of the parameters α and β are 1.97×10^{-5} and 0.959 respectively. Using these two parameters, we can determine the linking behavior in terms of the probability of linking to a node with degree (d) using Equation 17.

[Fig. 8 about here.]

[Fig. 9 about here.]

3.3.3 Results from Approach 3: Multinomial Logit Choice Model

In this section, we apply the multinomial logit model to deduce the node-level utility given the assumption that the node-level decision follows the form of Equation (15). In the multinomial logit choice model, each existing node is an alternative. The size of the network is large (e.g., the number of nodes in the network of Jan. 5^{th}, 2004 is 16301). Hence, to reduce the computational burden, the size of the choice set is reduced by grouping the nodes with same degree together as one alternative. The resulting probability is that of a newly added node selecting a given group representing a particular degree. An individual node's probability of getting a connection can then be obtained by assuming that all nodes within a group have the same probability.

In order to use the multinomial logit model, the first step is to identify the attributes to be considered in the systematic component of the utility function. We consider two aspects in the utility function: the node's degree (d_j) and the number of nodes with degree (n_j). Instead of using these parameters directly in the utility function, we use the natural logarithms of these parameters as the attributes of the nodes. Hence,

$$V_j = \beta_1 \ln(d_j) + \beta_2 \ln(n_j) \tag{18}$$

where β_1 is the parameter corresponding to degree d_j, β_2 is the parameter corresponding to group size n_j. This choice of the functional form is used because it results in a node-level behavior that can be directly estimated using existing multinomial logit algorithms. The parameters β_1 and β_2 denote the preferences of the decision-making nodes on degree and group size. Thus the utility function based on Equation (12) is:

$$u_j = \beta_1 \ln(d_j) + \beta_2 \ln(n_j) + \varepsilon_j \tag{19}$$

The resulting probability that the group j is chosen by node n in a network is:

$$P_n(j|C_J) = \frac{d_j^{\beta_1} n_j^{\beta_2}}{\sum_{i=1}^{J} d_i^{\beta_1} n_i^{\beta_2}} \tag{20}$$

[Fig. 10 about here.]

[Fig. 11 about here.]

The parameters β_1 and β_2 can be estimated by using the information from the observed network structure datasets. Thus the utility function in Equation (19) can be determined. We estimate the parameters β_1 and β_2 for all the 122 network datasets. The parameters are plotted against the network size in Figures 10 and 11. It is observed in Figure 10 that the parameter (β_1) has an average value of 0.672 for network size smaller than 21000 nodes (corresponding to the Internet network on Jan. 2, 2006), and an average value of 0.428 afterwards. This is verified through hypothesis tests on the slope, as discussed in the previous section. The p-values corresponding to $\{H_0^1 : S_{\beta_1} = 0$ vs. $H_1^1 : S_{\beta_1} \neq 0\}$ for network size, $J \leq 21000$ and $J > 21000$ are 0.03 and 0.11 respectively. We cannot reject the null hypothesis at a 2% level of significance. Hence, we claim that there has been no statistically significant change in slopes of β_1 within ranges $J \leq 21000$ and $J > 21000$. The parameter (β_2) follows a similar trend. The average value of β_2 for $J \leq 21000$ is 0.661 (p-value = 0.12) and for $J > 21000$, the average value of β_2 is 0.525 (p-value = 0.06).

3.4 Comparison of the Node-level Behaviors and Resulting Networks Using the Three Approaches

In this section, compare the node-level behaviors obtained by the three approaches and evaluate the generated network structures for a given time. We also discuss the generality, extensibility and computational capability of each approach.

3.4.1 Node-level Behaviors and Generated Networks

Figure 12 shows an example of the node-level linking behavior of the AS-level Internet network on Jan. 5^{th}, 2004, estimated using three different approaches. In Approach 1, the probability that a node is chosen is determined by Equation (4) with estimated additional attractiveness (A). In Approach 2, the node-level behavior is described by Equation (17), where the parameters α and β of the power function are estimated using the ordinary least square regression (OLS). In the third approach, the node-level behavior is obtained by Equation (20) and parameters β_1 and β_2 are estimated using the multinomial logit choice models. A comparison of the estimated linking behaviors is shown in Table 3.

[Fig. 12 about here.]

[Table 3 about here.]

Based on the node-level behavior models derived by using the three approaches, the network topology of Internet on Nov. 12^{th}, 2007 is simulated using the real network topology on Jan. 5^{th}, 2004 as the initial network. Since the A value (additional attractiveness) in Approach 1 has different trends using OLS estimation and maximum likelihood estimation, we

also use the A-value from both methods and compare the resulting networks. Figure 13 shows the degree distribution of the simulated network structure based on different node-level behaviors. It is observed that all the four degree distribution functions are close to each other. To quantify the differences between the four simulated degree distributions from the original network, the Kullback-Leibler (KL) divergence [32] measure is adopted. The KL divergence is non-negative and zero if the distributions match exactly. It is the expectation of the logarithmic difference between two probability distributions. The larger the value, the less likely it is that the two distributions are the same. The KL divergence is calculated based on the probability mass instead of the cumulative distribution. Figure 14 shows the logarithmic difference between the simulated distribution and the true distribution at each degree point. The KL divergence values are as follows:

1. GPA with OLS approach: 0.098
2. GPA with maximum-likelihood approach: 0.192
3. Statistical regression-based approach: 0.141
4. DCM based approach: 0.271

[Fig. 13 about here.]

[Fig. 14 about here.]

The results show that the degree distribution of the network, which is generated with the node behavior model estimated by Approach 1 with GPA model and OLS estimation, is more likely to match the degree distribution of the true Internet AS network compared with other three approaches. However, to compare and evaluate the three approaches, we also compare other commonly used network measures, shown in the Table 4, to evaluate the differences among the generated networks. As shown in Table 5, the number of nodes and edges added during each step are the same for all the simulated networks because the network formation process is the same and the difference is in the linking probability only. The average path lengths (APL) of the generated networks are close to the true value. Specifically, the APL in Approach 1 with OLS estimation is slightly lower than the true value, but the APL values with other approaches are slightly higher than the true network. We observe that the clustering coefficients of all the simulated networks are less than that of the true network. This can be explained as follows. High clustering coefficient in the AS-level Internet network results from a large number of peer-to-peer relationships between ASes. However, in the approaches used in this paper, the peer-to-peer mechanism is not explicitly included. We also observe a significant difference in the diameters of the true network and simulated networks. The diameters of the simulated networks are around 10, whereas the diameter of the true network is 17. The potential reason for this difference in the diameter is that the models in the three approaches do not account for the geographic aspects. In the real world, when a new customer AS joins the network, it prefers to purchase service from a nearby provider in order to minimize the linking and routing costs. As a summary, some differences are observed between the real Internet network and the generated networks. These differences are due to a) the estimation process itself, and b) assumptions made about the network formation process for the specific case study.

[Table 4 about here.]

[Table 5 about here.]

3.4.2 Generality and Extensibility

Out of the three approaches, the first approach has the advantage of being simple and easy to evaluate because of the direct function mapping between degree distribution and node-level attributes, A, i.e., the node's additional attractiveness. However, it is based on an assumed behavior model of preferential attachment, and can be applied only if the network degree distribution follows the power-law form. This is the major limitation for the first approach. Both the second and third approaches have potential generality to be used in other similar problems in complex networks, without placing such assumptions on the node-level behavior in advance.

The strategy of regarding the network evolution as a decision-making process is a promising approach to model the evolution of endogenous networks. It has the advantage of providing an explanatory framework for the relationship between the node-level preferences, node-level behaviors and the network structure. This is a fundamental aspect that the other two approaches fail to address. Additionally, it provides a framework to integrate existing decision-centric models, such as ABM and network formation game models, with the available network structure datasets. As we discussed before, the DCM proposed in this paper does not account for other attributes of ASes, such as economic, traffic, geographic attributes. Existing ABMs for the Internet have included these attributes. Thus, with the DCM, it is possible to setup a more realistic model for reconstructing the Internet topology if information about these attributes is available.

3.4.3 Computational Capability

A barrier in implementing the decision-centric approach is the computational burden. Since the number of alternatives is large, the estimation problem becomes computationally expensive if each node is treated as an alternative. A potential approach to manage this complexity is to reduce the size of the choice set by grouping the nodes with similar characteristics (e.g., degree) into one group, and treating each group as an alternative as shown in the case study. This is only valid if each node within a group is equivalent. Therefore, if the network size is small (e.g. hundreds or thousands of nodes) and explanations are needed for interpreting the nodes' behaviors, the third approach is a better choice, otherwise Approaches 1 and 2 are better.

4 CLOSING COMMENTS

The approaches discussed in this paper have potential applications in the area of complex systems analysis and design in three ways. First, the proposed approaches are general enough to be used for other complex systems where network structure datasets at different steps in the evolution process are available. Second, gaining an understanding of the underlying decisions can help in creating better models to describe real-world systems. Third, obtaining a network's node-level behavior can help in guiding the evolution of complex systems. By using the decision models of the nodes, the future evolution of the network can be predicted, assuming that the preferences of the nodes would remain the same. The models provide information about how the network is expected to evolve, and how the future structure and performance can be directed by influencing node-level decision models. Hence, we can influence the future network performance. The contribution of this paper does not only help researchers gain a better understanding about how a real-world complex system evolves, but also provides a comparison of different approaches for generating network topologies using datasets.

In summary, we show three approaches for estimating the node-level behaviors and preferences from network structure datasets. An illustrative example, AS level Internet network, is used to show how the approaches can be applied to estimate the AS-level behavior and to reconstruct the Internet topology with the estimated AS-level behaviors. Future work would focus on relaxing some of the assumptions made in the models. For example, in this paper, we only consider the addition of nodes in the network evolution. However, in reality, removal of nodes and link re-direction also occurs during network evolution. Thus, consideration of decisions to remove nodes and to redirect links would be an essential aspect for further improvement of the models. Additionally, other attributes of the nodes that may influence the node's utility will be further investigated. For example, in the Internet case study, economic factors such as the link creation costs and profits, geographic location, type of AS, etc. play an important role in the creation of new links. Therefore these variables will be investigated in the future to refine the discrete choice based model. Like any other data-driven modeling activity, the proposed approaches cannot be utilized if the system is new or if it is impractical to collect network structure data. In such cases, the alternative approach is to interview the decision makers to generate decision models. In scenarios where partial information about network structure is available, it can be used in conjunction with surveys and interviews to build a more reliable model. Such issues are avenues for potential future work. Finally, by utilizing theories from other disciplines such as economics, possible incentives can then be designed and introduced into the network to achieve desired network properties or performance.

5 ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support from the National Science Foundation grants 1265622 and 1201114, and China Scholarship Council award No. 2010628122. The authors also thank Jyotishka Datta for helping with the statistical analysis.

References

[1] Pahl, G., and Beitz, W., 1996. *Engineering Design: A Systematic Approach*, 2nd ed. Springer, London.

[2] Buede, D. M., 2000. *The Engineering Design of Systems: Models and Methods*. John Wiley and Sons, Inc., New York, N.Y.

[3] NASA, 2007. *NASA Systems Engineering Handbook (NASA/SP-2007-6105 Rev1)*. National Aeronautics and Space Administration, Washington, DC.

[4] Hawthorne, B. D., and Panchal, J. H., 2012. "Policy design for sustainable energy systems considering multiple objectives and incomplete preferences". In 2012 ASME International Design Engineering and Technical Conferences (Design Automation Conference), Chicago, IL, Paper number: DETC2012-70426.

[5] Sha, Z., and Panchal, J., 2013. "Towards the design of complex evolving networks with high robustness and resilience". In Procedia Computer Science, Proceedings of the 2013 Conference on Systems Engineering Research (CSER), Vol. 16, pp. 522–531.

[6] Barabasi, A. L., and Albert, R., 1999. "Emergence of scaling in random networks". *Science,* ***286***(5439), pp. 509–512.

[7] Albert, R., and Barabasi, A. L., 2002. "Statistical mechanics of complex networks". *Review of Modern Physics,* ***74***(1), pp. 47–97.
[8] Dorogovtsev, S. N., and Mendes, J. F. F., 2002. "Evolution of networks". *Advances in Physics,* ***51***(4), pp. 1079–1187.
[9] Faloutsos, M., Faloutsos, P., and Faloutsos, C., 1999. "On power-law relationships of the internet topology". In SIGCOMM '99: Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, pp. 251 – 262.
[10] Barabasi, A. L., Albert, R., and Jeong, H., 2000. "Scale-free characteristics of random networks: the topology of the world-wide web". *Physical A: Statistical Mechanisms and Its Applications,* ***281***(1-4), pp. 69–77.
[11] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabsi, A. L., 2000. "The large-scale organization of metabolic networks". *Nature,* ***407***(6804), pp. 651–653.
[12] Tangmunarunkit, H., Govindan, R., Jamin, S., Shenker, S., and Willinger, W., 2002. "Network topology generators: Degree-based vs. structural". In SIGCOMM '02 Proceedings of the 2002 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, pp. 147–159.
[13] Hogg, R. V., Craig, A., and W., M. J., 2004. *Introduction to Mathematical Statistics*, 6th ed. Prentice Hall.
[14] Clauset, A., Shalizi, C. R., and Newman, M. E. J., 2009. "Power-law distributions in empirical data". *SIAM Review,* ***51***(4), pp. 661–703.
[15] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P., 2007. *Numerical Recipes: The Art of Scientific Computation*, 3rd ed. Cambridge University Press, New York, NY.
[16] Clauset, A., Newman, M. E. J., and Moore, C., 2007. "Finding community structure in very large networks". *Physical Review E,* ***70***(6), p. 066111.
[17] Ben-Akiva, M., and Lerman, S. R., 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, MA.
[18] Williams, H. C. W. L., 1977. "On the formation of travel demand models and economic evaluation measures of user benefit". *Environment and Planning,* ***9***(3), pp. 285–344.
[19] Chen, W., Hoyle, C., and Wassenaar, H. J., 2013. *Decision-Based Design: Integrating Consumer Preferences into Engineering Design*. Springer.
[20] Train, K., 2009. *Discrete Choice Methods with Simulation*, second ed. Cambridge University Press, New York, NY.
[21] Viton, P. A., 2012. *Discrete-Choice Logit Models with R*. http://facweb.knowlton.ohio-state.edu/pviton/courses2/crp5700/5700-mlogit.pdf.
[22] Croissant, Y., 2012. *Estimation of multinomial logit models in R: The mlogit Packages*. http://cran.r-project.org/web/packages/mlogit/vignettes/mlogit.pdf.
[23] R Development Core Team, 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, http://www.r-project.org/.
[24] Watts, D. J., and Strogatz, S., 1998. "Collective dynamics of "small-world" networks". *Nature,* ***393***, pp. 440–442.
[25] Freeman, L., 1977. "A set of measures of centrality based on betweenness". *Sociometry,* ***40***(1), pp. 35–41.
[26] Leguay, J., 2004. "An analysis on the internet topology". PhD thesis, Linkoping University, Sweden.
[27] Hawkinson, J., and Bates, T., 1996. *Guidelines for creation, selection, and registration of an Autonomous System (AS)*. Network Working Group, Online: http://tools.ietf.org/html/rfc1930.
[28] Chen, Q., Chang, H., Govindan, R., Jamin, S., Shenker, S. J., and Willinger, W., 2002. "The origin of power laws in internet topologies revisited". In INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, Vol. 2, pp. 608–617.
[29] CAIDA, 2013. *The Cooperative Association for Internet Data Analysis*. http://www.caida.org/home/.
[30] RouterView, 2013. *The RouterView Project*. http://www.routeviews.org/.
[31] Newman, M. E., 2003. "The structure and function of complex networks". *SIAM Review,* ***45***(2), pp. 167–256.
[32] Kullback, S., and Leibler, R., 1951. "On information and sufficiency". *Annals of Mathematical Statistics,* ***22***(1), pp. 79–86.

List of Figures

1 Five levels and the associated mappings in complex endogenously evolving networks 15
2 Complementary cumulative degree distribution of networks generated by generalized BA model with different A values 16
3 Complementary cumulative degree distribution of AS-level Internet on Jan 5^{th}, 2004 17
4 Exponent (γ) in the degree distribution vs. network size 18
5 Number of edges (E) vs network size (J). 19
6 The exponent (γ) of 122 AS-level Internet networks obtained with the Maximum Likelihood fitting with goodness-of-fit tests based on Kolmogorov-Smirnov (KS) statistic and likelihood ratio. 20
7 Node-level behavior for three network evolution steps 21
8 Network size vs. α in the node's decision model 22
9 Network size vs. β in the node's decision model 23
10 Network size vs. parameter β_1 in the node's decision model 24
11 Network size vs. parameter β_2 in the node's decision model 25
12 Comparison of the node-level behavior of AS-level Internet on Jan. 5^{th}, 2004 deduced by three approaches . 26
13 Comparison of complementary cumulative degree distribution between the real network and simulated network with three approaches 27
14 KL divergence on the degree distribution of simulated networks 28

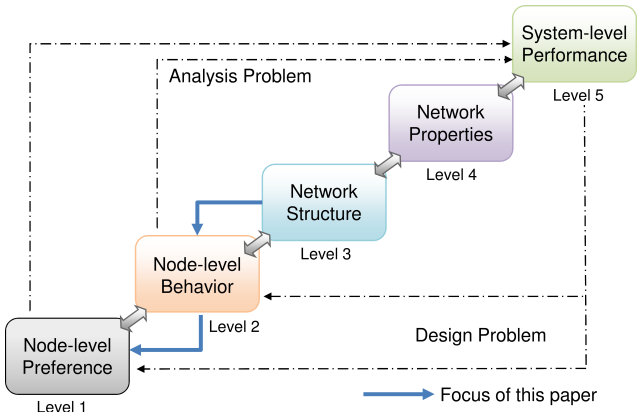

Fig. 1. Five levels and the associated mappings in complex endogenously evolving networks

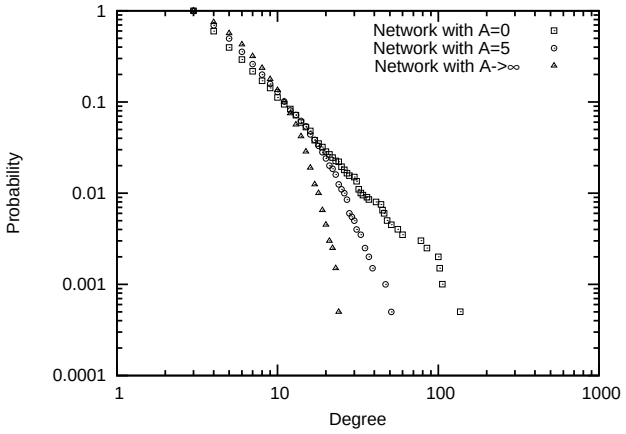

Fig. 2. Complementary cumulative degree distribution of networks generated by generalized BA model with different A values

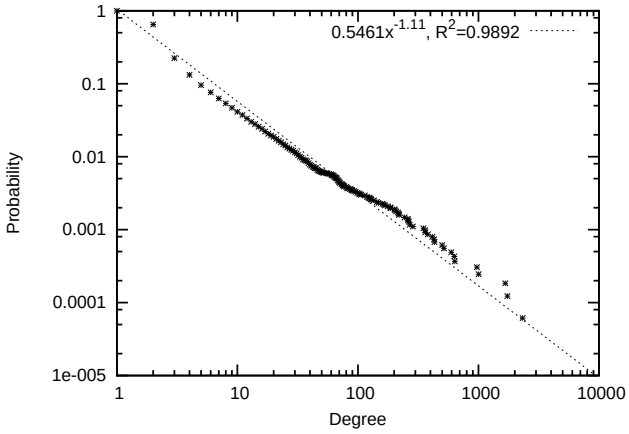

Fig. 3. Complementary cumulative degree distribution of AS-level Internet on Jan 5^{th}, 2004

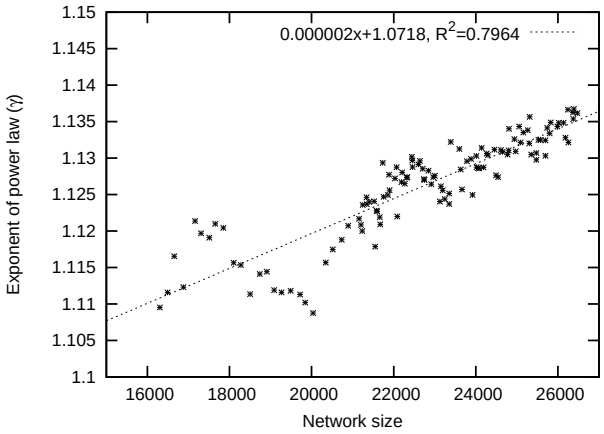

Fig. 4. Exponent (γ) in the degree distribution vs. network size

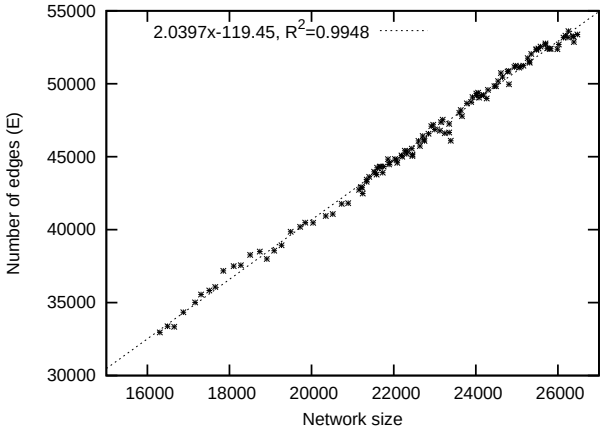

Fig. 5. Number of edges (E) vs network size (J).

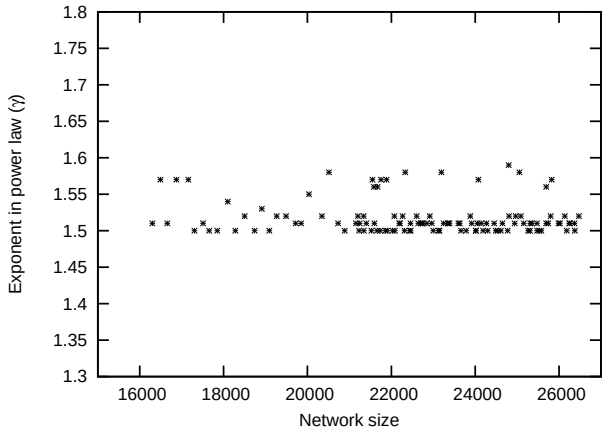

Fig. 6. The exponent (γ) of 122 AS-level Internet networks obtained with the Maximum Likelihood fitting with goodness-of-fit tests based on Kolmogorov-Smirnov (KS) statistic and likelihood ratio.

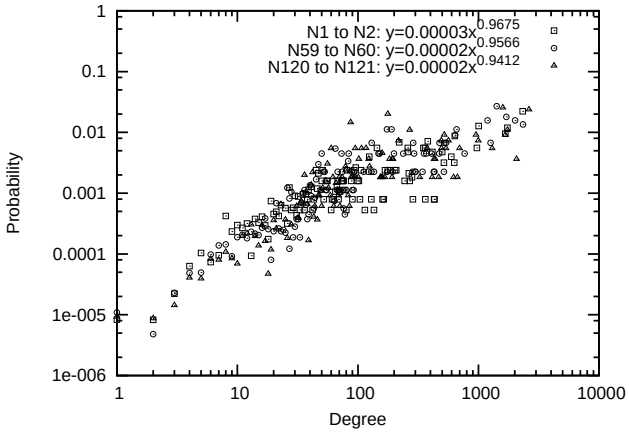

Fig. 7. Node-level behavior for three network evolution steps

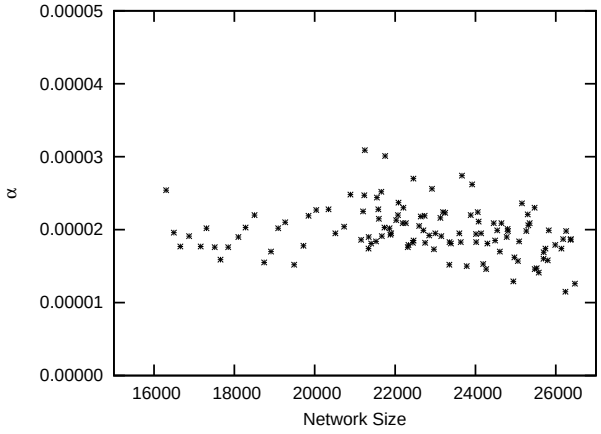

Fig. 8. Network size vs. α in the node's decision model

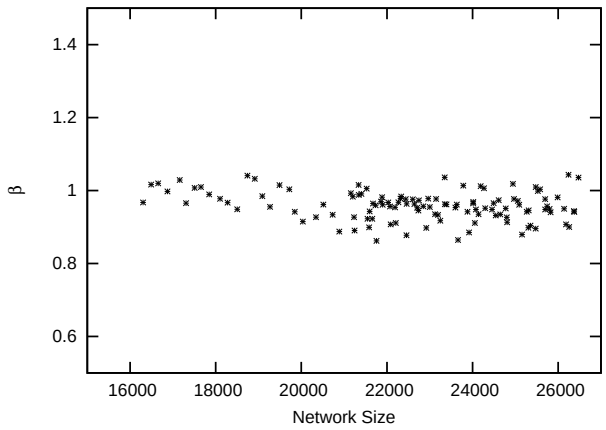

Fig. 9. Network size vs. β in the node's decision model

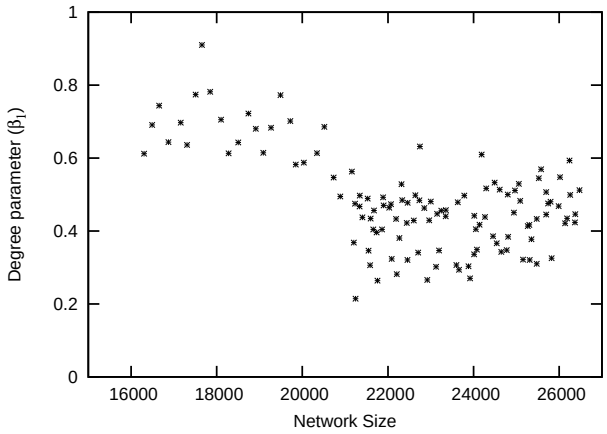

Fig. 10. Network size vs. parameter β_1 in the node's decision model

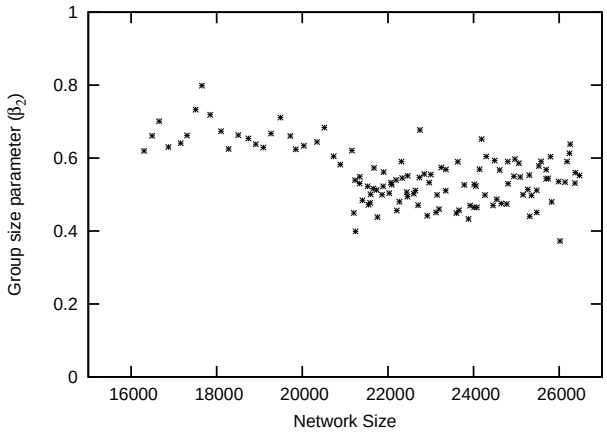

Fig. 11. Network size vs. parameter β_2 in the node's decision model

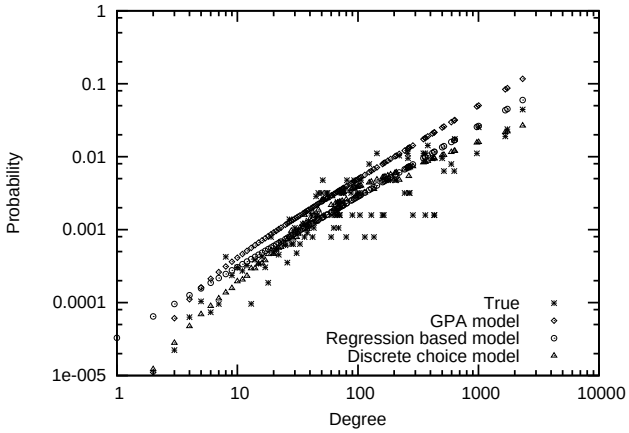

Fig. 12. Comparison of the node-level behavior of AS-level Internet on Jan. 5^{th}, 2004 deduced by three approaches

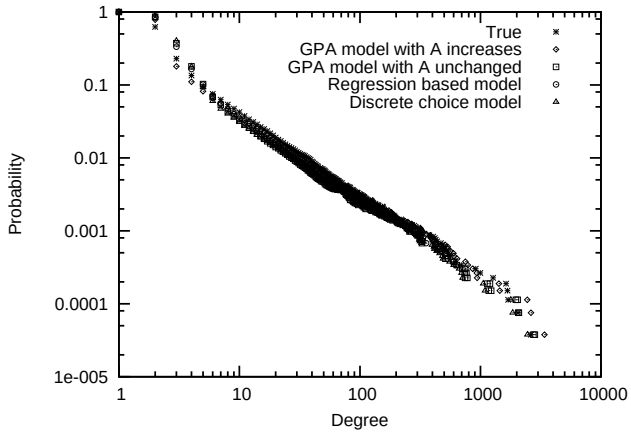

Fig. 13. Comparison of complementary cumulative degree distribution between the real network and simulated network with three approaches

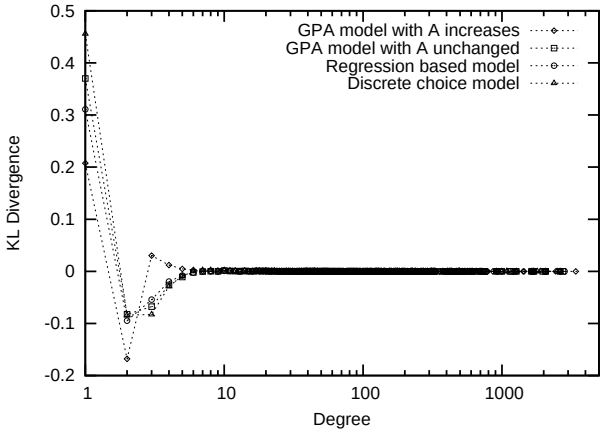

Fig. 14. KL divergence on the degree distribution of simulated networks

List of Tables

1 Distinction between hierarchically designed systems and complex evolutionary systems 30
2 An overview of the approaches presented in this paper 31
3 Linking probability comparison 32
4 Metrics for evaluation 33
5 Comparison of metrics in simulated networks 34

Table 1. Distinction between hierarchically designed systems and complex evolutionary systems

	Hierarchically Designed Systems	**Complex Evolutionary Systems**
Defining characteristics	Set of interacting components forming an integrated whole by designers	System evolves as a result of decisions and behaviors of individual self-directed entities
Local entities	No decision-making ability	Self-directed, generally selfish
Design variables	Components' attributes such as dimensions, material, etc.	Incentives to local entities and interacting mechanisms
Design strategy	Top-down hierarchical design	Bottom-up evolutionary design
Examples of design problems	Design of transportation network layout, traditional power grid assignment and design, optimizing flows on networks etc.	Traffic mechanism design, protocol design for Internet, policy design for green energy, incentive design etc.

Table 2. An overview of the approaches presented in this paper

	Generalized preferential attachment (Section 2.1)	**Statistical Regression-based approach** (Section 2.2)	**Multinomial-logit choice model** (Sectio 2.3)
Inputs	Single Instance of Network	Consecutive instances of the network	Consecutive instances of the model
Outputs	Linking probability	Linking probability	Linking probability and node's utility
Approach	Continuum theory	Regression	Discrete choice
Behavior model	$P_j(t) = \frac{G_j(t)}{\sum_{i=1}^{n} V_i(t)} d_j(t) + \frac{A_j(t)}{\sum_{i=1}^{n} V_i(t)}$	Best fitting model. Power model is adopted in this paper. $p_j = \alpha d_j^{\beta}$	$P(j\|C_J) = \frac{e^{V_{jn}}}{\sum_{i=1}^{J} e^{V_{in}}}$ Utility function: $V_j = \beta_{0j} + \beta_{1j} d_j$
Parameters in the model	G: Node's fitness; A: node's additional attractiveness	α: coefficient; β: exponent	Vector β_j that captures a node's preference
Estimation technique	Function mapping	Ordinary least square	Maximum likelihood

Table 3. Linking probability comparison

Approaches	Node-level Behavior Models
Approach 1 - GPA Model with OLS estimation	$P_j = \frac{d_j+A}{\sum_{i=1}^{J-1}(d_i+A)}$ where A changes over time according to $5 \times 10^{-6}J - 1.86$, and J is the number of nodes in the network at time t
Approach 1 - GPA Model with with Maximum likelihood estimation	$P_j = \frac{d_j-0.96}{\sum_{i=1}^{J-1}(d_i-0.96)}$
Approach 2 - Regression-based Model	$P_j = 1.97 \times 10^{-5} d_j^{0.959}$
Approach 3 - DCM	$P_n(j\|C_J) = \frac{d_j^{\beta_1} n_j^{\beta_2}}{\sum_{i=1}^{J} d_i^{\beta_1} n_i^{\beta_2}}$ $\beta_1 = 0.672$ and $\beta_2 = 0.661$ before Jan. 2^{nd}, 2006. $\beta_1 = 0.428$ and $\beta_2 = 0.525$ afterwards.

Table 4. Metrics for evaluation

Metrics	Relationship to Internet
Average Path Length (APL)	Related to routing efficiency
Cluster Coefficient	Related to peering structure and route resilience
Diameter	Related to the span of Internet

Table 5. Comparison of metrics in simulated networks

	# of Nodes	# of Edges	Cluster Coeffi.	APL	Diameter
True Network	26475	53381	0.208	3.876	17
Approach 1 with OLS	26475	53303	0.199	3.649	10
Approach 1 with Maximum likelihood	26475	53303	0.108	3.973	9
Approach 2	26475	53303	0.118	3.903	10
Approach 3	26475	53303	0.104	4.043	10